

Modified Linear Projection for Large Spatial Data Sets

Toshihiro Hirano*

February 25, 2014

Abstract

Recent developments in engineering techniques for spatial data collection such as geographic information systems have resulted in an increasing need for methods to analyze large spatial data sets. These sorts of data sets can be found in various fields of the natural and social sciences. However, model fitting and spatial prediction using these large spatial data sets are impractically time-consuming, because of the necessary matrix inversions. Various methods have been developed to deal with this problem, including the reduced rank approach and the sparse matrix approximation. In this paper, we propose a modification to an existing reduced rank approach to capture both the large- and small-scale spatial variations effectively. We have used simulated examples and empirical data analysis to demonstrate that our proposed approach consistently performs well when compared with other methods. In particular, the performance of our new method does not depend on the dependence properties of the spatial covariance functions.

Key words : Covariance tapering, Gaussian process, Geostatistics, Markov chain Monte Carlo, Reduced rank approximation, Stochastic matrix approximation

1 Introduction

Spatial data set analysis has been attracting an increasing amount of attention from various fields such as environmental science and economics, but is often impractical for large spatial data sets. This is because model fitting and spatial prediction in a Gaussian process model involve the inversion of an $n \times n$ covariance matrix for a data set of size n , which typically requires $O(n^3)$ operations. There is a rich literature regarding efficient computation for large spatial data sets (e.g., Stein et al. 2004; Fuentes 2007; Matsuda and Yajima 2009; Lindgren et al. 2011).

*Graduate School of Economics, The University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113-0033 Japan. Email: nennekog@hotmail.co.jp

In this paper, we consider two recently developed approaches that appeal as general purpose methodologies. The first approach is based on a reduced rank approximation of the underlying process. Cressie and Johannesson (2008) considered fixed rank approaches for kriging in large spatial data sets. Banerjee et al. (2008) proposed a predictive process that used a finite number of knots and Finley et al. (2009) corrected a bias in the predictive process. Recently, Banerjee et al. (2013) developed a linear projection approach in the literature of a Gaussian process regression (see, e.g., Rasmussen and Williams 2006). As in Banerjee et al. (2013), this is an extension of the predictive process and has the advantage of avoiding the complicated knot selection problem.

The second approach is a covariance tapering proposed by Furrer et al. (2006). The basic idea of the covariance tapering is to reduce a spatial covariance function to zero beyond some range by multiplying the true spatial covariance function by a positive definite, compactly supported function. Then, the resulting covariance matrix is sufficiently sparse to achieve computational efficiency with the sparse matrix algorithm (see, e.g., Davis 2006). Furrer et al. (2006) proved the asymptotic efficiency of the best linear unbiased predictor (BLUP) using the covariance tapering which is called the tapered BLUP for the original BLUP. Kaufman et al. (2008) applied the covariance tapering to the log-likelihood function and showed that the estimators maximizing the tapered approximation of the log-likelihood are strongly consistent. Hirano and Yajima (2013) investigated the asymptotic property of the prediction by the covariance tapering in a transformed random field.

Sang and Huang (2012) demonstrated that the predictive process fails to accurately approximate the small-scale dependence structure and the covariance tapering fails at long-scales. They proposed a combination of the predictive process and the covariance tapering, which is called a full scale approximation. Our paper confirms the same approximation property for the linear projection as the predictive process through some examples. We will show only one example in this paper. To deal with this problem, we propose a modified linear projection using the covariance tapering based on the work of Sang and Huang (2012).

The main contributions of this paper are to propose a linear projection using a modification by the compactly supported correlation function and investigate a theoretical justification. Furthermore, we have used simulated examples and empirical analysis based on air dose rate data to demonstrate that our proposed method works well when compared with the linear projection and the covariance tapering, regardless of the strength of spatial correlation and nonstationarity. Our work can be regarded as an extension of Banerjee et al. (2008), Finley et al. (2009), Sang and Huang (2012), and Banerjee et al. (2013). However, we have only focused on the linear projection approach proposed by Banerjee et al. (2013)

for comparison purposes because it outperformed the predictive process in the simulations and empirical studies of Banerjee et al. (2013) and the methods proposed by Finley et al. (2009) and Sang and Huang (2012) are a modification of the predictive process by the indicator function and the compactly supported correlation function respectively. Moreover, the selection of tuning parameters and many indices such as the estimation accuracy, prediction accuracy, and computational time make fair comparisons difficult.

The remainder of this paper is organized as follows. We introduce a linear regression model and Bayesian analysis for spatial data sets in Section 2. In Section 3, we review the linear projection approach and its algorithm. Section 4 presents our proposed modified linear projection. In Section 5, we present the results of computer experiments that compared the performance of our method with the linear projection and covariance tapering. Section 6 provides an empirical analysis based on the air dose rate in Chiba prefecture of eastern Japan. Our conclusions and future studies are discussed in Section 7. Technical proofs of the propositions are given in the Appendix.

2 Linear regression model and Bayesian analysis for spatial data sets

For $\mathbf{s} = (s_1, \dots, s_d)' (\in D \subset \mathbb{R}^d)$, consider the linear regression model of the form

$$Y(\mathbf{s}) = \mathbf{x}(\mathbf{s})' \boldsymbol{\beta} + W(\mathbf{s}) + \epsilon(\mathbf{s}), \quad (1)$$

where $Y(\mathbf{s})$ is a dependent variable at a location \mathbf{s} , $\mathbf{x}(\mathbf{s}) = (x_1(\mathbf{s}), \dots, x_p(\mathbf{s}))'$ is a p -vector of nonstochastic regressors, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ is a vector of unknown regression coefficients, and the prime denotes the transposition. The residual of this regression is decomposed into a zero-mean Gaussian process $W(\mathbf{s})$ with a valid covariance function $C_W(\mathbf{s}, \mathbf{s}^*) = \text{cov}(W(\mathbf{s}), W(\mathbf{s}^*))$ ($\mathbf{s}, \mathbf{s}^* \in D$) and $\epsilon(\mathbf{s})$ which is a zero-mean independent process following a normal distribution with a variance τ^2 for any location \mathbf{s} . $\epsilon(\mathbf{s})$ represents the possibility of measurement error and/or microscale variability and is often referred to as a nugget effect (see, e.g., Cressie 1993). It is assumed that $\{W(\mathbf{s})\}$ and $\{\epsilon(\mathbf{s})\}$ are independent. We specify that $C_W(\mathbf{s}, \mathbf{s}^*) = \sigma^2 \rho_W(\mathbf{s}, \mathbf{s}^*; \boldsymbol{\theta})$ where $\sigma^2 = \text{var}(W(\mathbf{s}))$, ρ_W is a correlation function of the spatial process $W(\mathbf{s})$, and $\boldsymbol{\theta}$ is a vector of correlation parameters.

Along with a $p \times 1$ vector of spatially referenced regressors $\mathbf{x}(\mathbf{s})$, we observe the dependent variable $Y(\mathbf{s})$ at given sampling locations $\mathbf{s}_1, \dots, \mathbf{s}_n \in D$. Denote $\mathbf{Y} = (Y(\mathbf{s}_1), \dots, Y(\mathbf{s}_n))'$

and $\Omega = (\boldsymbol{\beta}, \tau^2, \sigma^2, \boldsymbol{\theta})$. Then, the probability density function of \mathbf{Y} is

$$f(\mathbf{Y}|\Omega) = (2\pi)^{-\frac{n}{2}} |\Sigma_W + \tau^2 \mathbf{I}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\mathbf{Y} - X\boldsymbol{\beta})' (\Sigma_W + \tau^2 \mathbf{I})^{-1} (\mathbf{Y} - X\boldsymbol{\beta}) \right\},$$

where $(\Sigma_W)_{ij} = \sigma^2 \rho_W(\mathbf{s}_i, \mathbf{s}_j; \boldsymbol{\theta})$ ($i, j = 1, \dots, n$), \mathbf{I} is a $n \times n$ identity matrix, and $X = (\mathbf{x}(\mathbf{s}_1), \dots, \mathbf{x}(\mathbf{s}_n))'$. The goal is to estimate the parameters $\Omega = (\boldsymbol{\beta}, \tau^2, \sigma^2, \boldsymbol{\theta})$ and predict $Y(\mathbf{s}_0)$ at an unobserved location $\mathbf{s}_0 \in D$ based on \mathbf{Y} . Note that $\mathbf{x}(\mathbf{s}_0)$ is observed.

In this paper, we take a Bayesian approach and use a simulation method, namely, the Markov chain Monte Carlo (MCMC) method to generate samples from the posterior distribution and conduct statistical inference with respect to the model parameters. The Bayesian approach assigns prior distributions to $\Omega = (\boldsymbol{\beta}, \tau^2, \sigma^2, \boldsymbol{\theta})$ and the MCMC method is used to draw samples of the model parameters from the posterior distribution

$$\pi(\Omega|\mathbf{Y}) \propto f(\mathbf{Y}|\Omega) \pi(\boldsymbol{\beta}) \pi(\tau^2) \pi(\sigma^2) \pi(\boldsymbol{\theta}).$$

For prior distributions of $\boldsymbol{\beta}$, σ^2 , and τ^2 , we assume

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}_\beta, \Sigma_\beta), \quad \tau^2 \sim \mathcal{IG}(a_1, b_1), \quad \sigma^2 \sim \mathcal{IG}(a_2, b_2), \quad (2)$$

where $\mathcal{N}(\boldsymbol{\mu}_\beta, \Sigma_\beta)$ and $\mathcal{IG}(a_i, b_i)$ ($i = 1, 2$) respectively denote the multivariate normal distribution and inverse gamma distributions with probability density functions

$$\begin{aligned} \pi(\boldsymbol{\beta}) &\propto |\Sigma_\beta|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\mu}_\beta)' \Sigma_\beta^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu}_\beta) \right\}, \quad \pi(\tau^2) \propto (\tau^2)^{-(a_1+1)} \exp \left(-\frac{b_1}{\tau^2} \right), \\ \pi(\sigma^2) &\propto (\sigma^2)^{-(a_2+1)} \exp \left(-\frac{b_2}{\sigma^2} \right). \end{aligned}$$

Since the prior specifications for $\boldsymbol{\theta}$ will depend on the choice of correlation function $\rho_W(\mathbf{s}, \mathbf{s}^*; \boldsymbol{\theta})$, details of the estimation are discussed in Sections 5 and 6. We implement the MCMC algorithm in four stages:

1. Generate $\boldsymbol{\beta}|\tau^2, \sigma^2, \boldsymbol{\theta}, \mathbf{Y}$.
2. Generate $\tau^2|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \mathbf{Y}$.
3. Generate $\sigma^2|\boldsymbol{\beta}, \tau^2, \boldsymbol{\theta}, \mathbf{Y}$.
4. Generate $\boldsymbol{\theta}|\boldsymbol{\beta}, \tau^2, \sigma^2, \mathbf{Y}$.

Note that Σ_W depends on σ^2 and $\boldsymbol{\theta}$.

Generation of $\boldsymbol{\beta}$.

The conditional posterior probability density function of $\boldsymbol{\beta}$ is

$$\pi(\boldsymbol{\beta}|\tau^2, \sigma^2, \boldsymbol{\theta}, \mathbf{Y}) \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\beta}|\cdot}, \Sigma_{\boldsymbol{\beta}|\cdot}),$$

where

$$\Sigma_{\boldsymbol{\beta}|\cdot} = \left\{ \Sigma_{\boldsymbol{\beta}}^{-1} + X' (\Sigma_W + \tau^2 \mathbf{I})^{-1} X \right\}^{-1}$$

and

$$\boldsymbol{\mu}_{\boldsymbol{\beta}|\cdot} = \Sigma_{\boldsymbol{\beta}|\cdot} \left\{ \Sigma_{\boldsymbol{\beta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\beta}} + X' (\Sigma_W + \tau^2 \mathbf{I})^{-1} \mathbf{Y} \right\}.$$

τ^2 and σ^2 are updated using Metropolis steps (see, e.g., Gelman et al. 2004). Random-walk Metropolis steps with normal proposals are typically adopted.

Generation of τ^2 .

Given the current value τ^2 , propose a candidate $\tau^{2*} = \tau^2 + z_1$, $z_1 \sim \mathcal{N}(0, \sigma_1^2)$ and accept it with probability

$$\min \left[\frac{|\Sigma_W + \tau^{2*} \mathbf{I}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\mathbf{Y} - X\boldsymbol{\beta})' (\Sigma_W + \tau^{2*} \mathbf{I})^{-1} (\mathbf{Y} - X\boldsymbol{\beta}) - \frac{b_1}{\tau^{2*}} \right\} (\tau^{2*})^{-(a_1+1)}}{|\Sigma_W + \tau^2 \mathbf{I}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\mathbf{Y} - X\boldsymbol{\beta})' (\Sigma_W + \tau^2 \mathbf{I})^{-1} (\mathbf{Y} - X\boldsymbol{\beta}) - \frac{b_1}{\tau^2} \right\} (\tau^2)^{-(a_1+1)}}, 1 \right].$$

Generation of σ^2 .

Given the current value σ^2 , propose a candidate $\sigma^{2*} = \sigma^2 + z_2$, $z_2 \sim \mathcal{N}(0, \sigma_2^2)$ and accept it with probability

$$\min \left[\frac{|\Sigma_W^* + \tau^2 \mathbf{I}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\mathbf{Y} - X\boldsymbol{\beta})' (\Sigma_W^* + \tau^2 \mathbf{I})^{-1} (\mathbf{Y} - X\boldsymbol{\beta}) - \frac{b_2}{\sigma^{2*}} \right\} (\sigma^{2*})^{-(a_2+1)}}{|\Sigma_W + \tau^2 \mathbf{I}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\mathbf{Y} - X\boldsymbol{\beta})' (\Sigma_W + \tau^2 \mathbf{I})^{-1} (\mathbf{Y} - X\boldsymbol{\beta}) - \frac{b_2}{\sigma^2} \right\} (\sigma^2)^{-(a_2+1)}}, 1 \right],$$

where $(\Sigma_W^*)_{ij} = \sigma^{2*} \rho_W(\mathbf{s}_i, \mathbf{s}_j; \boldsymbol{\theta})$ ($i, j = 1, \dots, n$). The tuning parameters σ_1^2 and σ_2^2 are chosen such that the average of the acceptance rates in each iteration is approximately 40%. As we have previously mentioned, the generation of $\boldsymbol{\theta}$ will be discussed in subsequent sections.

It is computationally expensive to calculate the determinant and inverse of the $n \times n$ matrix $\Sigma_W + \tau^2 \mathbf{I}$ for large spatial data sets. In particular, the inverse matrix calculation requires $O(n^3)$ operations. For each sampling procedure, we must calculate the determinant and inverse of the $n \times n$ matrix $\Sigma_W + \tau^2 \mathbf{I}$. Thus, the computational complexity of the above MCMC algorithm is challenging for large spatial data sets because a large number of samples from the posterior distribution are usually needed.

The Bayesian prediction is to obtain the predictive distribution

$$\pi(Y(\mathbf{s}_0)|\mathbf{Y}) = \int \pi(Y(\mathbf{s}_0)|\mathbf{Y}, \Omega) \pi(\Omega|\mathbf{Y}) d\Omega.$$

For a given Ω ,

$$\pi(Y(\mathbf{s}_0)|\mathbf{Y}, \Omega) \sim \mathcal{N} \left(\mathbf{x}(\mathbf{s}_0)' \boldsymbol{\beta} + \mathbf{c}'_{Y, \mathbf{s}_0} (\Sigma_W + \tau^2 \mathbf{I})^{-1} (\mathbf{Y} - X \boldsymbol{\beta}), \sigma^2 + \tau^2 - \mathbf{c}'_{Y, \mathbf{s}_0} (\Sigma_W + \tau^2 \mathbf{I})^{-1} \mathbf{c}_{Y, \mathbf{s}_0} \right),$$

where $\mathbf{c}_{Y, \mathbf{s}_0} = (\text{cov}(Y(\mathbf{s}_0), Y(\mathbf{s}_1)), \dots, \text{cov}(Y(\mathbf{s}_0), Y(\mathbf{s}_n)))'$. The predictive distribution is sampled by composition, drawing $Y^{(l)}(\mathbf{s}_0) \sim \pi(Y(\mathbf{s}_0)|\mathbf{Y}, \Omega^{(l)})$ for each $\Omega^{(l)}$ ($l = 1, \dots, L$) where $\Omega^{(l)}$ is the l th sample from the posterior distribution $\pi(\Omega|\mathbf{Y})$ and L is the total number of samples given in the MCMC algorithm. The mean squared prediction error (MSPE) is computed using

$$\frac{1}{M} \sum_{m=1}^M \left(Y(\mathbf{s}_{0,m}) - \frac{1}{L} \sum_{l=1}^L Y^{(l)}(\mathbf{s}_{0,m}) \right)^2,$$

where $Y(\mathbf{s}_{0,m})$ ($m = 1, \dots, M$) and $\sum_{l=1}^L Y^{(l)}(\mathbf{s}_{0,m})/L$ denote the test data sets and the sample analogue of the mean of the predictive distribution respectively. Since sampling from $\pi(Y(\mathbf{s}_0)|\mathbf{Y}, \Omega^{(l)})$ also involves the inverse of the $n \times n$ matrix $\Sigma_W + \tau^2 \mathbf{I}$, the computation becomes a more formidable one for large spatial data sets.

Finally, we compare the computational efficiency of some existing approximation methods using the deviance information criterion (DIC) (Spiegelhalter et al. 2002). It is used as a Bayesian measure of fit or adequacy and is defined as

$$DIC = E_{\Omega|\mathbf{Y}}[D(\Omega)] + p_D,$$

where $D(\Omega) = -2 \log f(\mathbf{Y}|\Omega)$ and $p_D = E_{\Omega|\mathbf{Y}}[D(\Omega)] - D(E_{\Omega|\mathbf{Y}}[\Omega])$. $E_{\Omega|\mathbf{Y}}[\cdot]$ represents the expectation under the posterior distribution $\pi(\Omega|\mathbf{Y})$. To compute $E_{\Omega|\mathbf{Y}}[D(\Omega)]$ and $E_{\Omega|\mathbf{Y}}[\Omega]$, we use the sample analogues

$$\frac{1}{L} \sum_{l=1}^L D(\Omega^{(l)}) \quad \text{and} \quad \frac{1}{L} \sum_{l=1}^L \Omega^{(l)}.$$

DIC also includes the inversion of the $n \times n$ matrix.

3 Linear projection approach

In this section, we review the linear projection approach proposed by Banerjee et al. (2013). This method was developed to efficiently compute the Gaussian process regression. However, it can be applied to the efficient computation of the Bayesian analysis for large spatial data

sets and the linear projection approach is regarded as an extension of predictive process models (Banerjee et al. 2008; Finley et al. 2009) as in Banerjee et al. (2013).

As a first step, for $\mathbf{s} \in D$, define

$$W_{approx}(\mathbf{s}) = E[W(\mathbf{s})|\Phi\mathbf{W}] = \mathbf{c}'_{W,\mathbf{s}}\Phi'(\Phi\Sigma_W\Phi')^{-1}\Phi\mathbf{W},$$

where $\mathbf{W} = (W(\mathbf{s}_1), \dots, W(\mathbf{s}_n))'$, $\mathbf{c}_{W,\mathbf{s}} = (\text{cov}(W(\mathbf{s}), W(\mathbf{s}_1)), \dots, \text{cov}(W(\mathbf{s}), W(\mathbf{s}_n)))' = (C_W(\mathbf{s}, \mathbf{s}_1), \dots, C_W(\mathbf{s}, \mathbf{s}_n))'$, and Φ is an $m \times n$ matrix with full row-rank ($m \leq n$) and a row-norm equal to unity to avoid scale problems. In this case, for $\mathbf{s}, \mathbf{s}^* \in D$,

$$C_{approx}(\mathbf{s}, \mathbf{s}^*) = \text{cov}(W_{approx}(\mathbf{s}), W_{approx}(\mathbf{s}^*)) = \mathbf{c}'_{W,\mathbf{s}}\Phi'(\Phi\Sigma_W\Phi')^{-1}\Phi\mathbf{c}_{W,\mathbf{s}^*}.$$

Since C_{approx} underestimates the variance of $W(\mathbf{s})$ from $E[(W(\mathbf{s}) - E[W(\mathbf{s})|\Phi\mathbf{W}])^2] = C_W(\mathbf{s}, \mathbf{s}) - \mathbf{c}'_{W,\mathbf{s}}\Phi'(\Phi\Sigma_W\Phi')^{-1}\Phi\mathbf{c}_{W,\mathbf{s}} \geq 0$, Banerjee et al. (2013) defined

$$C_{lp}(\mathbf{s}, \mathbf{s}^*) = C_{approx}(\mathbf{s}, \mathbf{s}^*) + \delta(\mathbf{s}, \mathbf{s}^*) \{C_W(\mathbf{s}, \mathbf{s}^*) - C_{approx}(\mathbf{s}, \mathbf{s}^*)\}, \quad (3)$$

where $\delta(\mathbf{s}, \mathbf{s}^*)$ is 1 if $\mathbf{s} = \mathbf{s}^*$, otherwise 0. This modification is based on Finley et al. (2009). Let $\{W_{lp}(\mathbf{s})\}$ be a zero-mean Gaussian random field with the covariance function C_{lp} . Σ_{approx} , Σ_{diag} , and Σ_{lp} denote the $n \times n$ covariance matrices with the (i, j) -th element of $C_{approx}(\mathbf{s}_i, \mathbf{s}_j)$, $\delta(\mathbf{s}_i, \mathbf{s}_j)(C_W(\mathbf{s}_i, \mathbf{s}_j) - C_{approx}(\mathbf{s}_i, \mathbf{s}_j))$, and $C_{lp}(\mathbf{s}_i, \mathbf{s}_j)$ respectively. These matrix expressions are given by

$$\begin{aligned} \Sigma_{approx} &= \Sigma_W\Phi'(\Phi\Sigma_W\Phi')^{-1}\Phi\Sigma_W, \\ \Sigma_{diag} &= \left\{ \Sigma_W - \Sigma_W\Phi'(\Phi\Sigma_W\Phi')^{-1}\Phi\Sigma_W \right\} \circ \mathbf{I}, \end{aligned}$$

and

$$\Sigma_{lp} = \Sigma_{approx} + \Sigma_{diag},$$

where the notation ' \circ ' refers to the Hadamard product. Now, we replace $W(\mathbf{s})$ in (1) with $W_{lp}(\mathbf{s})$. Consequently, the covariance matrix of \mathbf{Y} changes from $\Sigma_W + \tau^2\mathbf{I}$ to $\Sigma_{lp} + \tau^2\mathbf{I} = \Sigma_W\Phi'(\Phi\Sigma_W\Phi')^{-1}\Phi\Sigma_W + \Sigma_{diag} + \tau^2\mathbf{I}$ and the inverse matrix and determinant of $\Sigma_W + \tau^2\mathbf{I}$ in the Bayesian inference and prediction of Section 2 become those of $\Sigma_{lp} + \tau^2\mathbf{I}$. Using Harville (1997), we obtain

$$\begin{aligned} \left\{ \Sigma_W\Phi'(\Phi\Sigma_W\Phi')^{-1}\Phi\Sigma_W + \Sigma_{diag} + \tau^2\mathbf{I} \right\}^{-1} &= (\Sigma_{diag} + \tau^2\mathbf{I})^{-1} - (\Sigma_{diag} + \tau^2\mathbf{I})^{-1}\Sigma_W\Phi' \\ &\times \left\{ \Phi\Sigma_W\Phi' + \Phi\Sigma_W(\Sigma_{diag} + \tau^2\mathbf{I})^{-1}\Sigma_W\Phi' \right\}^{-1}\Phi\Sigma_W(\Sigma_{diag} + \tau^2\mathbf{I})^{-1}. \end{aligned} \quad (4)$$

Similarly, from Harville (1997), the determinant can be calculated using

$$\begin{aligned} \left| \Sigma_W \Phi' (\Phi \Sigma_W \Phi')^{-1} \Phi \Sigma_W + \Sigma_{diag} + \tau^2 \mathbf{I} \right| &= \left| \Sigma_{diag} + \tau^2 \mathbf{I} \right| \left| \Phi \Sigma_W \Phi' \right|^{-1} \\ &\times \left| \Phi \Sigma_W \Phi' + \Phi \Sigma_W (\Sigma_{diag} + \tau^2 \mathbf{I})^{-1} \Sigma_W \Phi' \right|. \end{aligned} \quad (5)$$

The right-hand sides of (4) and (5) include the inversion and determinant of the $n \times n$ diagonal matrix $\Sigma_{diag} + \tau^2 \mathbf{I}$ and the $m \times m$ matrices, so that it is faster to conduct the Bayesian inference and prediction. If Φ is an $m \times n$ submatrix of an $n \times n$ permutation matrix, we obtain a predictive process whose knots are an m -dimensional subset of $\{\mathbf{s}_1, \dots, \mathbf{s}_n\}$. Therefore, the linear projection is an extension of predictive process models. Additionally, the linear projection approach avoids the knot selection problem of the predictive process.

Next, we explain the selection of Φ using the stochastic matrix approximation technique in Banerjee et al. (2013). It follows from Schmidt's approximation theorem (Stewart 1993; page 563) that $U'_m = \arg \min_{\Phi} \|\Sigma_W - \Sigma_W \Phi' (\Phi \Sigma_W \Phi')^{-1} \Phi \Sigma_W\|_F$ for fixed m where $\|\cdot\|_F$ denotes the Frobenius norm for a matrix and U_m is the $n \times m$ matrix whose i th column vector is the eigenvector corresponding to the i th eigenvalue of Σ_W in descending order of magnitude ($i = 1, \dots, n$). However, the derivation of eigenvalues and eigenvectors of Σ_W involves $O(n^3)$ computations (Golub and Van Loan 1996). From $U_m U'_m \Sigma_W = \Sigma_W U_m (U'_m \Sigma_W U_m)^{-1} U'_m \Sigma_W$, Banerjee et al. (2013) proposed the following algorithm to find Φ by diminishing $\|\Sigma_W - \Phi' \Phi \Sigma_W\|_F$ for any target error level on the basis of the appropriate modification of Algorithm 4.2 of Halko et al. (2011).

Algorithm (Banerjee et al. 2013). Given a target error $\epsilon > 0$ and $r \in \mathbb{N}$, find the $m \times n$ matrix Φ that satisfies $\|\Sigma_W - \Phi' \Phi \Sigma_W\|_F < \epsilon$ with probability $1 - n/10^r$.

Step 1. Initialize $j = 0$ and $\Phi^{(0)} = \mathbf{I}$ (the $0 \times n$ empty matrix).

Step 2. Draw r length- n random vectors $\boldsymbol{\omega}^{(1)}, \dots, \boldsymbol{\omega}^{(r)}$ with independent entries from $\mathcal{N}(0, 1)$.

Step 3. Compute $\boldsymbol{\kappa}^{(i)} = \Sigma_W \boldsymbol{\omega}^{(i)}$ for $i = 1, \dots, r$.

Step 4. Check if $\max_{i=1, \dots, r} (\|\boldsymbol{\kappa}^{(i+j)}\|) < \{(\pi/2)^{1/2} \epsilon\}/10$. If it holds, go to Step 11. Otherwise go to Step 5.

Step 5. Recompute $j = j + 1$, $\boldsymbol{\kappa}^{(j)} = [\mathbf{I} - \{\Phi^{(j-1)}\}' \Phi^{(j-1)}] \boldsymbol{\kappa}^{(j)}$, and $\boldsymbol{\phi}^{(j)} = \boldsymbol{\kappa}^{(j)} / \|\boldsymbol{\kappa}^{(j)}\|$.

Step 6. Set $\Phi^{(j)} = [\{\Phi^{(j-1)}\}' \quad \boldsymbol{\phi}^{(j)}]'$.

Step 7. Draw a length- n random vector $\boldsymbol{\omega}^{(j+r)}$ with independent entries from $\mathcal{N}(0, 1)$.

Step 8. Compute $\boldsymbol{\kappa}^{(j+r)} = [\mathbf{I} - \{\Phi^{(j)}\}' \Phi^{(j)}] \Sigma_W \boldsymbol{\omega}^{(j+r)}$.

Step 9. Recompute $\boldsymbol{\kappa}^{(i)} = \boldsymbol{\kappa}^{(i)} - \boldsymbol{\phi}^{(j)} \{(\boldsymbol{\phi}^{(j)})' \boldsymbol{\kappa}^{(i)}\}$ for $i = (j + 1), \dots, (j + r - 1)$.

Step 10. Go back to the target error check in Step 4.

Step 11. If $j = 0$, output $\Phi = \{\boldsymbol{\kappa}^{(1)} / \|\boldsymbol{\kappa}^{(1)}\|\}'$; else output $\Phi = \Phi^{(j)}$.

Here, $\|\cdot\|$ denotes the Euclidean norm. Step 5 is not essential, but it ensures better stability when $\kappa^{(j)}$ becomes very small (see Halko et al. 2011). Step 6 combines the matrix and the vector. Banerjee et al. (2013) evaluated the linear projection approach using simulations and empirical examples and demonstrated that it achieved the better performance efficiently than the predictive process of Banerjee et al. (2008).

4 Modified linear projection

As previously mentioned, the linear projection approach is not related to the knot selection problem unlike the predictive process. However, similarly to the predictive process, the linear projection approach is inaccurate when approximating local or small-scale dependences of the true covariance function C_W . In contrast, it is effective for the predictive process and the linear projection to capture long-scale spatial variations of C_W .

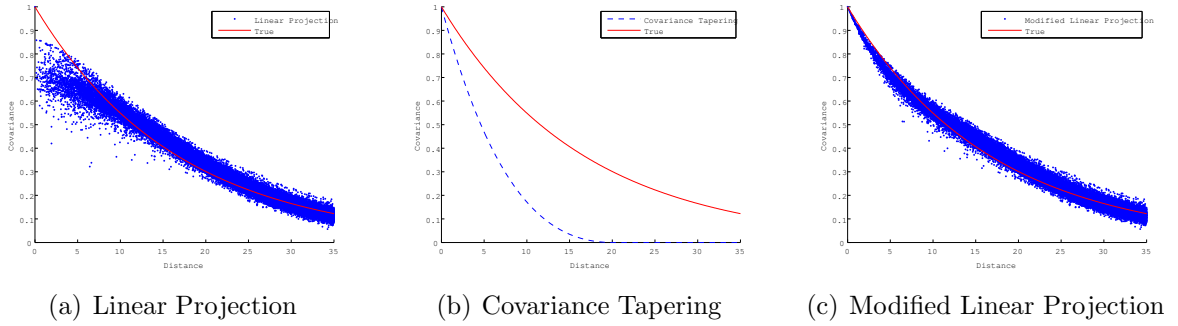


Figure 1: Exponential covariance function $C_W(\mathbf{s}, \mathbf{s}^*) = \exp(-0.06\|\mathbf{s} - \mathbf{s}^*\|)$ (solid line) and three approximations. The true covariance matrix is generated using 500 random locations in $[0, 100] \times [0, 100]$. (a) Linear projection approach with $\epsilon = 200$ and $r = 4$ (points). $m = 26$ was selected. (b) Covariance tapering using the spherical covariance function with $\gamma = 20$ (dotted line). (c) Modified linear projection using the linear projection with $\epsilon = 200$ and $r = 4$ and the spherical covariance function with $\gamma = 20$ (points).

Figure 1(a) shows a typical example to demonstrate that the improvement such as (3) is insufficient for modifying the approximation of small-scale dependence in the original covariance function. Using the linear projection, we obtain similar results for other covariance functions through some simulations (e.g., Gaussian covariance function and the Cauchy family (Gneiting and Schlather 2004)).

Our proposed approach is a modification of the linear projection approach by the covariance tapering, which is based on the idea of the full scale approximation in Sang and Huang (2012). Before introducing our new approach, we review the covariance tapering which gen-

erates the sparse matrix approximation from the compactly supported correlation function and achieves the computational efficiency for analyzing large spatial data sets.

Let $K_\gamma(x)$ ($x \geq 0$ and $\gamma > 0$) be a compactly supported correlation function with $K_\gamma(0) = 1$ and $K_\gamma(x) = 0$ for $x \geq \gamma$. $K_\gamma(x)$ is called the taper function with a taper range γ . Some compactly supported correlation functions have been developed (see, e.g., Wendland 1995). For example, there are the spherical covariance function

$$K_\gamma(x) = \left(1 - \frac{x}{\gamma}\right)_+^2 \left(1 + \frac{x}{2\gamma}\right)$$

and

$$K_\gamma(x) = \left(1 - \frac{x}{\gamma}\right)_+^6 \left(1 + 6\frac{x}{\gamma} + \frac{35x^2}{3\gamma^2}\right). \quad (6)$$

Now, consider the product of the original covariance function and the taper function, that is

$$C_{ct}(\mathbf{s}, \mathbf{s}^*) = C_W(\mathbf{s}, \mathbf{s}^*) K_\gamma(\|\mathbf{s} - \mathbf{s}^*\|).$$

Let $\{W_{ct}(\mathbf{s})\}$ be a zero-mean Gaussian random field with the covariance function C_{ct} and replace $W(\mathbf{s})$ in (1) with $W_{ct}(\mathbf{s})$. Then, $\Sigma_W + \tau^2 \mathbf{I}$ in the Bayesian inference and prediction of Section 2 becomes $\Sigma_W \circ \Sigma_{taper} + \tau^2 \mathbf{I}$ where $(\Sigma_{taper})_{ij} = K_\gamma(\|\mathbf{s}_i - \mathbf{s}_j\|)$ ($i, j = 1, \dots, n$). The resulting matrix $\Sigma_W \circ \Sigma_{taper} + \tau^2 \mathbf{I}$ has many zero elements and is called a sparse matrix, so that we can use sparse matrix algorithms to efficiently handle the inverse matrix and determinant.

From the definition of the covariance tapering, small-scale spatial dependence is well approximated, but large-scale dependence may not be appropriately accounted for (see Figure 1(b)). We introduce a modified linear projection approach to the covariance function of the original spatial process. It allows for efficient computations when using large spatial data sets. Define

$$C_{mlp}(\mathbf{s}, \mathbf{s}^*) = C_{approx}(\mathbf{s}, \mathbf{s}^*) + K_\gamma(\|\mathbf{s} - \mathbf{s}^*\|) \{C_W(\mathbf{s}, \mathbf{s}^*) - C_{approx}(\mathbf{s}, \mathbf{s}^*)\}.$$

Now, let Σ_{sparse} and Σ_{mlp} be the $n \times n$ Gram matrices with respect to $\mathbf{s}_1, \dots, \mathbf{s}_n$ for $K_\gamma(C_W - C_{approx})$ and C_{mlp} respectively. These matrix expressions are given by

$$\Sigma_{sparse} = \left\{ \Sigma_W - \Sigma_W \Phi' (\Phi \Sigma_W \Phi')^{-1} \Phi \Sigma_W \right\} \circ \Sigma_{taper} \quad \text{and} \quad \Sigma_{mlp} = \Sigma_{approx} + \Sigma_{sparse}.$$

The following proposition states the associated theoretical properties, which are used for the valid expansion of the inversion and the determinant.

Proposition 1

- (a) $\Phi \Sigma_W \Phi'$ is positive definite.
- (b) $\Sigma_{sparse} + \tau^2 \mathbf{I}$ is positive definite.
- (c) $\Sigma_{mlp} + \tau^2 \mathbf{I}$ is positive definite.
- (d) $\Phi \Sigma_W \Phi' + \Phi \Sigma_W (\Sigma_{sparse} + \tau^2 \mathbf{I})^{-1} \Sigma_W \Phi'$ is positive definite.

In the proof of Proposition 1(c), we prove that Σ_{mlp} is positive semidefinite. An additional assumption on $\{W(\mathbf{s})\}$ yields the positive definiteness of Σ_{mlp} , but its proof is omitted for brevity. Consider a zero-mean Gaussian random field $\{W_{mlp}(\mathbf{s})\}$ with the covariance function C_{mlp} and replace $W(\mathbf{s})$ in (1) with $W_{mlp}(\mathbf{s})$. Consequently, $(\Sigma_W + \tau^2 \mathbf{I})^{-1}$ and $|\Sigma_W + \tau^2 \mathbf{I}|$ in the Bayesian inference and prediction become the inverse matrix and determinant of $\Sigma_{mlp} + \tau^2 \mathbf{I} = \Sigma_W \Phi' (\Phi \Sigma_W \Phi')^{-1} \Phi \Sigma_W + \Sigma_{sparse} + \tau^2 \mathbf{I}$ respectively. Σ_{mlp} includes the original covariance matrix Σ_W because $\Sigma_W = \Sigma_{mlp}$ if $m = n$. Similarly to the linear projection case, from Harville (1997) and Proposition 1, we can obtain

$$\begin{aligned} \left\{ \Sigma_W \Phi' (\Phi \Sigma_W \Phi')^{-1} \Phi \Sigma_W + \Sigma_{sparse} + \tau^2 \mathbf{I} \right\}^{-1} &= (\Sigma_{sparse} + \tau^2 \mathbf{I})^{-1} - (\Sigma_{sparse} + \tau^2 \mathbf{I})^{-1} \Sigma_W \Phi' \\ &\quad \times \left\{ \Phi \Sigma_W \Phi' + \Phi \Sigma_W (\Sigma_{sparse} + \tau^2 \mathbf{I})^{-1} \Sigma_W \Phi' \right\}^{-1} \Phi \Sigma_W (\Sigma_{sparse} + \tau^2 \mathbf{I})^{-1} \end{aligned} \quad (7)$$

and

$$\begin{aligned} \left| \Sigma_W \Phi' (\Phi \Sigma_W \Phi')^{-1} \Phi \Sigma_W + \Sigma_{sparse} + \tau^2 \mathbf{I} \right| &= \left| \Sigma_{sparse} + \tau^2 \mathbf{I} \right| \left| \Phi \Sigma_W \Phi' \right|^{-1} \\ &\quad \times \left| \Phi \Sigma_W \Phi' + \Phi \Sigma_W (\Sigma_{sparse} + \tau^2 \mathbf{I})^{-1} \Sigma_W \Phi' \right|. \end{aligned} \quad (8)$$

Now, we can treat the inverse matrix and determinant more quickly because (7) and (8) include the $n \times n$ sparse matrix $\Sigma_{sparse} + \tau^2 \mathbf{I}$ and $m \times m$ matrices. Figure 1(c) describes the good fitting of the modified linear projection to the original covariance function because the new approach uses the linear projection to capture large-scale spatial variations and the covariance tapering to capture small-scale local variations that are unexplained by the linear projection. In fact, the following proposition shows that the modified linear projection is superior to the linear projection in a sense of the Frobenius norm.

Proposition 2

Suppose that $K_{\gamma_1}(x) \leq K_{\gamma_2}(x)$ for fixed $x \geq 0$. Then,

$$\|\Sigma_W - \Sigma_{approx}\|_F \geq \|\Sigma_W - \Sigma_{lp}\|_F \geq \|\Sigma_W - \Sigma_{mlp, \gamma_1}\|_F \geq \|\Sigma_W - \Sigma_{mlp, \gamma_2}\|_F,$$

where $\Sigma_{mlp, \gamma_i} = \Sigma_{approx} + (\Sigma_W - \Sigma_{approx}) \circ \Sigma_{taper, \gamma_i}$ and $(\Sigma_{taper, \gamma_i})_{kl} = K_{\gamma_i}(\|\mathbf{s}_k - \mathbf{s}_l\|)$ for $k, l = 1, \dots, n$ and $i = 1, 2$.

From Proposition 2, the approximation of the modified linear projection is better than that of the linear projection with respect to the Frobenius norm. Furthermore, it follows that the condition of Proposition 2 is satisfied for the spherical covariance function and the taper function (6) when $\gamma_1 \leq \gamma_2$. Then, as the taper range γ increases, the approximation accuracy of the modified linear projection increases in a sense of the Frobenius norm. However, there is a trade-off between the magnitude of the taper range and the computational burden.

Finally, we now show that the approximation accuracy for the original covariance matrix controls the error in the probability density function of \mathbf{Y} . The next proposition is a corollary of Theorem 2 in Banerjee et al. (2013).

Proposition 3

Suppose that $\Sigma_*(A) = \Sigma_W \Phi' (\Phi \Sigma_W \Phi')^{-1} \Phi \Sigma_W + \{\Sigma_W - \Sigma_W \Phi' (\Phi \Sigma_W \Phi')^{-1} \Phi \Sigma_W\} \circ A$ and A is a $n \times n$ positive definite matrix. Let $f = \mathcal{N}(X\beta, \Sigma_W + \tau^2 \mathbf{I})$ be the probability density function of \mathbf{Y} under the original model and $f_* = \mathcal{N}(X\beta, \Sigma_*(A) + \tau^2 \mathbf{I})$ denotes its linear projection-type approximation. If $\|\Sigma_W - \Sigma_*(A)\|_F \leq \epsilon$ for sufficiently small $\epsilon > 0$, then

$$d_{KL}(f, f_*) \leq \frac{n}{2} \left\{ \frac{\epsilon}{\tau^2} - \log \left(1 - \frac{\epsilon}{\tau^2} \right) \right\},$$

where d_{KL} denotes the Kullback-Leibler divergence between probability density functions.

In other words, the Kullback-Leibler divergence is of the same order as the error in the approximation of the original covariance matrix in terms of the Frobenius norm. Since most of our derivation is a straightforward application of Banerjee et al. (2013) without a small gap, we omit the proof of Proposition 3. Note that $\Sigma_*(\mathbf{I}) = \Sigma_{lp}$ and $\Sigma_*(\Sigma_{taper}) = \Sigma_{mlp}$. From Proposition 3, the error between the original probability density function and that of the modified linear projection has the sharp bound compared to the linear projection because $\|\Sigma_W - \Sigma_{lp}\|_F \geq \|\Sigma_W - \Sigma_{mlp}\|_F$ in Proposition 2.

5 Illustrative examples using simulated data

This section illustrates our proposed method using simulated data and examines the effect of our modification using the compactly supported correlation function by comparing it with the linear projection and the covariance tapering. All computations were carried out using MATLAB functions `sparse`, `symamd`, and `chol` on Linux powered 2.50GHz Xeon processor with 64 Gbytes RAM. The convergence diagnostics and other posterior summarizations for MCMC were implemented by the R package `CODA` (Plummer et al. 2006). In our simulations and empirical study, the taper function (6) was used for the covariance tapering and modified linear projection.

First, we investigated the performance of the proposed method through a simple simulation. Let $D = [0, 100]^2$ be the sampling domain and 2000 locations were sampled from a uniform distribution over D . We randomly selected 1500 locations for the estimation of parameters and DIC, while the rest were used for the calculation of the MSPE. We employed the Matérn correlation function

$$\rho_W(\mathbf{s}, \mathbf{s}_*; \nu, \lambda) = \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{2\nu^{1/2}\|\mathbf{s} - \mathbf{s}_*\|}{\lambda} \right)^\nu J_\nu \left(\frac{2\nu^{1/2}\|\mathbf{s} - \mathbf{s}_*\|}{\lambda} \right), \quad \nu > 0, \lambda > 0,$$

where $\Gamma(\cdot)$ is the gamma function and J_ν is the modified Bessel function of the second kind of order ν (see Stein 1999). The spatial range parameter λ controls the decay in spatial correlation and the smoothness parameter ν can be interpreted as the degree of the smoothness of the random field. For example, if $\nu = 0.5$, the Matérn correlation function is

$$\rho_W(\mathbf{s}, \mathbf{s}_*; \lambda) = \exp \left(-\frac{\sqrt{2}\|\mathbf{s} - \mathbf{s}_*\|}{\lambda} \right). \quad (9)$$

This is called the exponential covariance function and is widely used in many applications. The data were simulated from the model (1) with $\boldsymbol{\beta} = \mathbf{0}$, the exponential covariance function with $\sigma^2 = 0.5$, $\lambda = \sqrt{2}/0.06$ and $\sqrt{2}/0.3$, and nugget variance $\tau^2 = 1$. σ^2 and τ^2 were the targets of the estimation and we assumed that the other parameters were known. When pairs of observations are more than 50 unit distant from each other in $\lambda = \sqrt{2}/0.06$, they have negligible (< 0.05) correlation. This distance is called the effective range of the random field and 50 unit represents the random field with the strong spatial correlation. Similarly, the effective range in $\lambda = \sqrt{2}/0.3$ is 10 unit and it has the weak spatial correlation. For prior distributions, we assumed that $a_1 = 1$, $b_1 = 0.1$, $a_2 = 0.8$, and $b_2 = 0.1$ in (2). We ran the second and third stages in the MCMC algorithm presented in Section 2 for 50000 iterations, discarding the first 500 samples as burn-in periods. We applied the linear projection with $\epsilon = 200$ and $r = 4$, the covariance tapering with $\gamma = 2.8$ and 20, and the modified linear projection with $\epsilon = 200$, $r = 4$, and $\gamma = 20$ in $\lambda = \sqrt{2}/0.06$. In $\lambda = \sqrt{2}/0.3$, the linear projection with $\epsilon = 150, 400$, and $r = 4$, the covariance tapering with $\gamma = 2.8$ and 10, and the modified linear projection with $\epsilon = 400$, $r = 4$, and $\gamma = 2.8$ were considered.

The inefficiency factor (IF) is defined as $1 + 2 \sum_{t=1}^{\infty} \rho(t)$ where $\rho(t)$ is the sample autocorrelation at lag t for the parameter of interest. This factor is used to measure how well the MCMC mixes (e.g., Chib 2001). The smaller the inefficiency factor becomes, the closer the MCMC sampling is to the uncorrelated one. The computational time of each approach is relative to the time taken in the full model, scaled to 1. These times include the calculation of MSPE, DIC, and Φ selected by the algorithm presented in Section 3. Additionally,

we described the rank of Φ required in the algorithm for ϵ and the sparsity of the matrix measured by the percentage of zero elements in the off-diagonal elements of Σ_{taper} for γ .

Table 1: Summary of results from the first simulation in $\lambda = \sqrt{2}/0.06$.

		τ^2	σ^2	MSPE	DIC	Relative time
True value		1	0.5	-	-	-
Original model	Mean	0.992	0.432	1.138	4472	1
	Stdev	0.043	0.090			
	95% interval	[0.911, 1.079]	[0.282, 0.633]			
	IF	7.537	21.753			
MLP ($\epsilon = 200, \gamma = 2.8$)	Mean	0.985	0.458	1.157	4479	0.38
	Stdev	0.047	0.111			
	95% interval	[0.895, 1.077]	[0.281, 0.708]			
	IF	12.919	35.117			
LP ($\epsilon = 200$)	Mean	0.987	0.451	1.158	4480	0.36
	Stdev	0.047	0.108			
	95% interval	[0.896, 1.079]	[0.276, 0.698]			
	IF	13.612	31.757			
CT ($\gamma = 2.8$)	Mean	0.769	0.596	1.399	4702	0.22
	Stdev	0.10	0.106			
	95% interval	[0.586, 0.979]	[0.388, 0.802]			
	IF	46.439	48.492			
CT ($\gamma = 20$)	Mean	0.971	0.342	1.158	4514	2.57
	Stdev	0.041	0.050			
	95% interval	[0.894, 1.056]	[0.255, 0.449]			
	IF	5.408	8.832			

MLP: modified linear projection; LP: linear projection; CT: covariance tapering. The required rank was 84 when $\epsilon = 200$. The sparsity was 0.24% when $\gamma = 2.8$ and 10.71% when $\gamma = 20$.

Table 2: Summary of results from the first simulation in $\lambda = \sqrt{2}/0.3$.

		τ^2	σ^2	MSPE	DIC	Relative time
True value		1	0.5	-	-	-
Original model	Mean	1.018	0.427	1.340	4717	1
	Stdev	0.061	0.074			
	95% interval	[0.90, 1.142]	[0.293, 0.582]			
	IF	14.042	19.599			
MLP ($\epsilon = 400, \gamma = 2.8$)	Mean	1.037	0.416	1.408	4754	0.35
	Stdev	0.095	0.116			
	95% interval	[0.836, 1.211]	[0.224, 0.678]			
	IF	47.139	55.115			
LP ($\epsilon = 400$)	Mean	1.119	0.319	1.433	4763	0.33
	Stdev	0.085	0.094			
	95% interval	[0.944, 1.277]	[0.161, 0.531]			
	IF	219.804	407.408			
CT ($\gamma = 2.8$)	Mean	0.763	0.665	1.444	4772	0.21
	Stdev	0.119	0.125			
	95% interval	[0.541, 1.011]	[0.412, 0.911]			
	IF	64.035	66.709			
LP ($\epsilon = 150$)	Mean	1.086	0.345	1.351	4724	1.2
	Stdev	0.056	0.058			
	95% interval	[0.981, 1.20]	[0.238, 0.466]			
	IF	37.746	149.150			
CT ($\gamma = 10$)	Mean	0.928	0.499	1.367	4727	0.79
	Stdev	0.066	0.075			
	95% interval	[0.80, 1.063]	[0.359, 0.651]			
	IF	16.551	20.983			

MLP: modified linear projection; LP: linear projection; CT: covariance tapering. The required rank was 87 when $\epsilon = 400$ and 510 when $\epsilon = 150$. The sparsity was 0.24% when $\gamma = 2.8$ and 2.94% when $\gamma = 10$.

Tables 1 and 2 display the Bayesian posterior sample means, standard deviations, and 95% credible intervals of the model parameters for each approach. Each approximation method required more time as ϵ decreased or γ increased, which often offset the computational efficiency. As shown in Table 1, the linear projection worked very well in the random field with the strong correlation. However, in the random field with the weak correlation,

Table 2 shows that the estimation of σ^2 using the linear projection was insufficient and the original model was superior to the linear projection even from a perspective of the calculation time. This is because the linear projection places a particular emphasis on fitting to the long-scale dependence. Since the covariance tapering has the property opposite to the linear projection, its performance was good except in the case where the linear projection was effective. Unlike the linear projection and the covariance tapering, the modified linear projection proposed in this paper performed well regardless of the magnitude of the spatial correlation. The modified linear projection with appropriate taper range γ improved the linear projection by adding a bit of time.

In the second simulation, we considered a nonstationary random field using a covariance function developed by Paciorek and Schervish (2006). The covariance function is

$$C_W(\mathbf{s}, \mathbf{s}_*) = \sigma^2 \frac{1}{2^{\nu-1}\Gamma(\nu)} |\Sigma_{D(\mathbf{s})}|^{\frac{1}{4}} |\Sigma_{D(\mathbf{s}_*)}|^{\frac{1}{4}} \left| \frac{\Sigma_{D(\mathbf{s})} + \Sigma_{D(\mathbf{s}_*)}}{2} \right|^{-\frac{1}{2}} \left\{ 2\sqrt{\nu d(\mathbf{s}, \mathbf{s}_*)} \right\}^\nu J_\nu \left\{ 2\sqrt{\nu d(\mathbf{s}, \mathbf{s}_*)} \right\}, \quad (10)$$

where $\Sigma_{D(\mathbf{s})}$ is a $d \times d$ positive definite matrix,

$$d(\mathbf{s}, \mathbf{s}_*) = (\mathbf{s} - \mathbf{s}_*)' \left(\frac{\Sigma_{D(\mathbf{s})} + \Sigma_{D(\mathbf{s}_*)}}{2} \right)^{-1} (\mathbf{s} - \mathbf{s}_*),$$

and $D(\mathbf{s})$ indicates the subregion which \mathbf{s} belongs to. In the second simulation, we considered the sampling domain $D = [0, 500]^2$ and partitioned the entire region D into two subregions $D_1 = [0, 250] \times [0, 500]$ and $D_2 = (250, 500] \times [0, 500]$. 1000 locations were sampled from a uniform distribution over D_i and split into 750 training sets and 250 test sets for $i = 1, 2$. As a result, we obtained 1500 locations for training and 500 locations for testing. For the nonstationary covariance function (10), the smoothness parameter ν was fixed to be 0.5 and we set $\Sigma_{D(\mathbf{s})} = \lambda_{D(\mathbf{s})}^2 \mathbf{I}$ where $D(\mathbf{s})$ is 1 if $\mathbf{s} \in D_1$, and 2 otherwise. This introduced the nonstationary random field that combined the stationary random field of the range parameter λ_1 over D_1 with that of the range parameter λ_2 over D_2 . In Section 3 of Paciorek and Schervish (2006), there is an example of spatial data sets with different stationary covariance structures in the eastern and western regions similar to the nonstationary random field in our simulation. The data sets were simulated using the spatial linear regression model (1) with $x_1(\mathbf{s}) = 1$, $x_2(\mathbf{s})$ generated from the standard normal distribution, $\boldsymbol{\beta} = (1, 2)'$, $\sigma^2 = 0.67$, $\tau^2 = 0.11$, $\lambda_1 = 1/0.08$, and $\lambda_2 = 1/0.3$. For prior distributions, we assumed that $\boldsymbol{\mu}_\beta = (0.959, 1.972)'$, $\Sigma_\beta = 1000\mathbf{I}$, $a_1 = 11$, $b_1 = 1.261$, $a_2 = 11$, and $b_2 = 6.305$. These hyperparameter choices were guided by the least squares estimator and an appropriate partition of the sample variance of its residual based on typical values in past empirical studies.

In order to bypass the computational burden of the selection of Φ at each iteration, a discrete uniform distribution with atoms $\{c_1, \dots, c_{t_j}\}$ was taken as the prior distribution of λ_j ($j = 1, 2$) because we can precompute Φ using the algorithm presented in Section 3 for the correlation matrix of $W(\mathbf{s})$ with each distinct value of $\lambda_1 \in \{c_1, \dots, c_{t_1}\}$ and $\lambda_2 \in \{c_1, \dots, c_{t_2}\}$ prior to implementing the MCMC procedure. This strategy was proposed in Section 4 of Banerjee et al. (2013). In addition, Wikle (2010) used the discrete uniform distribution as the prior distribution for the range parameter.

Generation of λ_1 .

For $i = 1, \dots, t_1$, the conditional posterior distribution of λ_1 is

$$\begin{aligned} P(\lambda_1 = c_i | \boldsymbol{\beta}, \tau^2, \sigma^2, \lambda_2, \mathbf{Y}) \\ = K_1 |\Sigma_W(\lambda_1 = c_i) + \tau^2 \mathbf{I}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\mathbf{Y} - X\boldsymbol{\beta})' (\Sigma_W(\lambda_1 = c_i) + \tau^2 \mathbf{I})^{-1} (\mathbf{Y} - X\boldsymbol{\beta}) \right\}, \end{aligned}$$

where $\Sigma_W(\lambda_1 = c_i)$ denotes Σ_W with $\lambda_1 = c_i$ and

$$K_1 = 1 / \sum_{i=1}^{t_1} |\Sigma_W(\lambda_1 = c_i) + \tau^2 \mathbf{I}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\mathbf{Y} - X\boldsymbol{\beta})' (\Sigma_W(\lambda_1 = c_i) + \tau^2 \mathbf{I})^{-1} (\mathbf{Y} - X\boldsymbol{\beta}) \right\}.$$

Generation of λ_2 .

Similarly, for $i = 1, \dots, t_2$, the conditional posterior distribution of λ_2 is

$$\begin{aligned} P(\lambda_2 = c_i | \boldsymbol{\beta}, \tau^2, \sigma^2, \lambda_1, \mathbf{Y}) \\ = K_2 |\Sigma_W(\lambda_2 = c_i) + \tau^2 \mathbf{I}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\mathbf{Y} - X\boldsymbol{\beta})' (\Sigma_W(\lambda_2 = c_i) + \tau^2 \mathbf{I})^{-1} (\mathbf{Y} - X\boldsymbol{\beta}) \right\}, \end{aligned}$$

where $\Sigma_W(\lambda_2 = c_i)$ denotes Σ_W with $\lambda_2 = c_i$ and

$$K_2 = 1 / \sum_{i=1}^{t_2} |\Sigma_W(\lambda_2 = c_i) + \tau^2 \mathbf{I}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\mathbf{Y} - X\boldsymbol{\beta})' (\Sigma_W(\lambda_2 = c_i) + \tau^2 \mathbf{I})^{-1} (\mathbf{Y} - X\boldsymbol{\beta}) \right\}.$$

For $j = 1, 2$, we set $c_i = 1/(0.02i)$ ($i = 1, \dots, t_j$) and $t_j = 25$ to choose a wide interval of the range parameters. The linear projection with $\epsilon = 350$ and $r = 5$, the covariance tapering with $\gamma = 12$, and the modified linear projection with $\epsilon = 350$, $r = 5$, and $\gamma = 12$ were applied. Using the MCMC algorithm described in Section 2, we sampled 8000 draws after the initial 300 samples were discarded as a burn-in period.

The results of the simulation are summarized in Table 3. For σ^2 and τ^2 , the linear projection showed high inefficiency factors and had discrepancies from the posterior means of the original model. This implies that the modification by the indicator function is not sufficient for small-scale variations. For β_1 and λ_1 , the 95% interval of the linear projection does not include the true value. The covariance tapering was highly computationally efficient, but the estimations of the range parameters were inaccurate. However, our proposed method modified these drawbacks and worked well compared to the linear projection and the covariance tapering. We obtained similar results with other settings, but these are not reported here.

Table 3: Summary of results from the second simulation.

	True value		Original	MLP ($\epsilon = 350, \gamma = 12$)	LP ($\epsilon = 350$)	CT ($\gamma = 12$)
β_1	1	Mean	0.949	0.952	0.933	0.962
		Stdev	0.029	0.030	0.031	0.023
		95% interval	[0.891, 1.004]	[0.892, 1.011]	[0.873, 0.994]	[0.916, 1.006]
		IF	1.425	5.641	2.530	1.379
β_2	2	Mean	1.980	1.972	1.970	1.970
		Stdev	0.020	0.021	0.022	0.026
		95% interval	[1.942, 2.019]	[1.930, 2.013]	[1.927, 2.012]	[1.929, 2.011]
		IF	1.0	1.0	1.0	1.0
τ^2	0.11	Mean	0.138	0.139	0.214	0.120
		Stdev	0.031	0.039	0.062	0.029
		95% interval	[0.086, 0.203]	[0.079, 0.223]	[0.099, 0.327]	[0.073, 0.185]
		IF	36.209	74.679	131.557	36.937
σ^2	0.67	Mean	0.628	0.643	0.589	0.629
		Stdev	0.041	0.049	0.080	0.039
		95% interval	[0.546, 0.706]	[0.546, 0.738]	[0.443, 0.743]	[0.548, 0.707]
		IF	26.513	40.110	129.658	26.337
λ_1	1/0.08	Mean	13.944	18.257	24.368	36.215
		Stdev	2.075	3.521	2.206	15.197
		95% interval	[12.50, 16.667]	[12.50, 25.0]	[16.667, 25.0]	[10.0, 50.0]
		IF	14.361	43.472	37.155	2.208
λ_2	1/0.3	Mean	2.714	3.045	3.341	4.493
		Stdev	0.523	1.089	1.020	2.075
		95% interval	[2.0, 3.846]	[2.174, 5.556]	[2.381, 5.556]	[2.083, 10.0]
		IF	1.131	1.269	2.952	1.263
MSPE	-		0.664	0.681	0.710	0.708
DIC	-		3701	3724	3746	3780
Relative time	-		1	0.54	0.52	0.36

MLP: modified linear projection; LP: linear projection; CT: covariance tapering. Average required rank and its 95% interval were 23.75 and [1, 94.50] when $\epsilon = 350$. The sparsity was 0.17% when $\gamma = 12$.

6 Empirical study

In this section, we discuss the results when we applied our proposed modified linear projection method to air dose rates in Chiba prefecture. This data are created based on the results of the vehicle-borne survey conducted by the Japanese Ministry of Education, Culture, Sports, Science and Technology from November 5 to December 10, 2012. This dataset consists of air dose rates (microsievert per hour) with longitudes, latitudes, and distances from the Fukushima Dai-ichi Nuclear Power Plant (NPP) (km) at 47470 sampling points and is obtained from the Environment Monitoring Database for the Distribution of Radioactive Substances Released by the TEPCO Fukushima Dai-ichi NPP Accident at <http://radb.jaea.go.jp/mapdb/en/>. These are spatio-temporal data because they were observed on irregularly spaced locations at discrete time points. However, we have considered the dataset to be spatial by assuming that the air dose rate trend does not fluctuate largely over a short period. To assume the Gaussian process over the whole region, we selected 5557 points inside the rectangular region $[139.920625, 140.103125] \times [35.25375, 35.424584]$. Figure 2 shows the logarithmic transformation of these spatial data sets. To understand the entire trend for the scattered observations, we attempted to make the prediction surface using representative points of the predictive distribution.

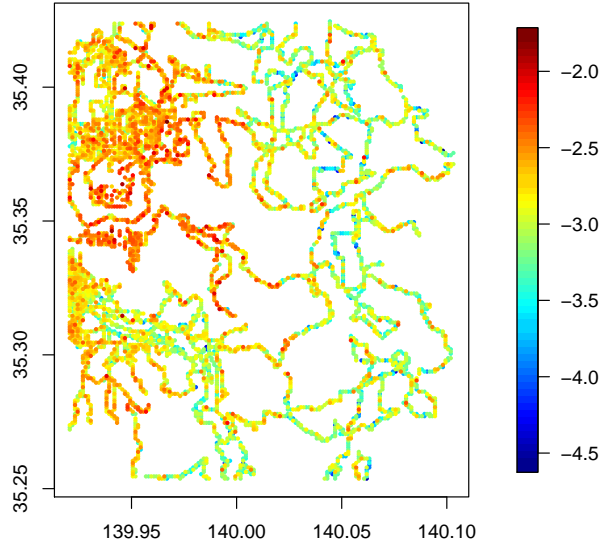


Figure 2: The logarithmic transformation of air dose rates at 5557 sampling points in Chiba prefecture.

We split 5557 data points into a training set of 5000 observations and a test set of 557 observations. To account for the mean component and maintain normality, we considered the spatial regression model (1) with the logarithmic transformation $Y(\mathbf{s})$ of the air dose rate, $x_1(\mathbf{s}) = 1$, and the distance from the Fukushima Dai-ichi NPP $x_2(\mathbf{s})$.

The MCMC algorithm was similar to the second simulation presented in Section 5. The hyperparameters of the prior distributions for $\boldsymbol{\beta} = (\beta_1, \beta_2)'$, σ^2 , and τ^2 were assumed to be $\boldsymbol{\mu}_{\boldsymbol{\beta}} = (-2.5849, -0.0012)'$, $\Sigma_{\boldsymbol{\beta}} = 1000\mathbf{I}$, $a_1 = 11$, $b_1 = 0.2121$, $a_2 = 11$, and $b_2 = 1.0607$ in the same way as the second simulation. Moreover, we conducted some trial runs using a training subset. The results led us to use the exponential covariance function (9) and the discrete uniform distribution with $c_i = 1/(0.01i)$ ($i = 1, \dots, 60$) for the prior distribution of λ . To implement the modified linear projection, we conducted a pilot analysis using training and test subsets for various choices of ϵ and γ . Weighing the trade-off between the prediction accuracy and run time, we determined that $\epsilon = 1200$, $r = 5$, and $\gamma = 0.5$ were appropriate choices. However, for $\epsilon = 1200$, the linear projection caused poor mixing of the sampling of λ because of the weak spatial correlation of the logarithmic transformation of the air dose rate. Hence, $\epsilon = 300$ and $r = 5$ were selected for the linear projection, which resulted in a longer run time than that of the modified linear projection.

To run the MCMC algorithm, we drew 4100 samples and discarded 100 samples as a burn-in period. The predictive surfaces were generated by considering the predictive distribution at 31×31 prediction points, which overlaid the sampling domain, and calculating the mean, 5%, and 95% quantiles of samples from the predictive distribution. The mean of the predictive distribution was used as the single point predictor and the 5% and 95% quantiles of the predictive distribution served as measures of uncertainty of the mean of the predictive distribution. Note that the calculation time for each method includes the generation of the predictive surfaces as well as the Bayesian estimation and Bayesian prediction.

Table 4: The result of the Bayesian analysis in the empirical data example.

		Stationary model	MLP ($\epsilon = 1200, \gamma = 0.5$)	LP ($\epsilon = 300$)	CT ($\gamma = 0.5$)
β_1	Mean	-3.325	-1.559	-3.295	-2.971
	Stdev	2.986	2.897	1.441	0.380
	95% interval	[-9.206, 2.615]	[-7.248, 3.974]	[-6.118, -0.434]	[-3.696, -2.218]
	IF	0.970	1.137	1.0	0.867
β_2	Mean	0.001	-0.006	0.001	0.0003
	Stdev	0.012	0.012	0.006	0.002
	95% interval	[-0.023, 0.025]	[-0.028, 0.018]	[-0.010, 0.013]	[-0.003, 0.003]
	IF	0.974	1.137	1.0	0.868
τ^2	Mean	0.070	0.055	0.068	0.049
	Stdev	0.002	0.002	0.002	0.002
	95% interval	[0.066, 0.073]	[0.051, 0.058]	[0.065, 0.072]	[0.046, 0.052]
	IF	11.872	15.055	13.0	9.452
σ^2	Mean	0.077	0.110	0.446	0.070
	Stdev	0.018	0.009	0.005	0.003
	95% interval	[0.050, 0.121]	[0.093, 0.129]	[0.036, 0.055]	[0.065, 0.076]
	IF	38.980	11.472	6.825	7.074
λ	Mean	4.645	4.187	1.735	35.601
	Stdev	1.594	0.308	0.182	31.961
	95% interval	[2.381, 8.333]	[3.704, 5.556]	[1.667, 2.041]	[4.762, 100.0]
	IF	32.214	4.20	4.912	1.202
MSPE		0.076	0.071	0.079	0.074
DIC		1703	1598	1747	2551
Relative time		1	0.51	0.59	0.34

MLP: modified linear projection; LP: linear projection; CT: covariance tapering. Average required rank and its 95% interval were 45.133 and [8, 77] when $\epsilon = 1200$ and 187.88 and [26, 335.50] when $\epsilon = 300$. The sparsity was 0.46% when $\gamma = 0.5$.

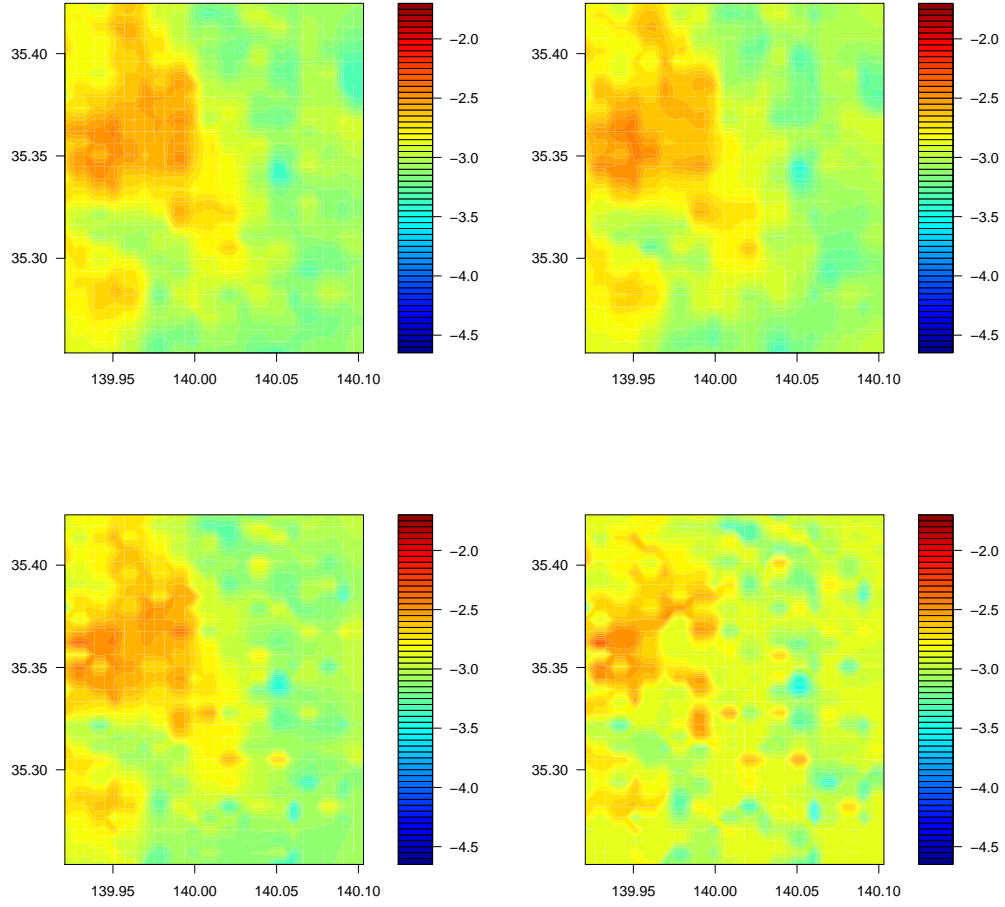


Figure 3: Mean of the predictive distribution. We adopt the stationary model and the linear projection for the left and right column in the top row respectively. The modified linear projection and the covariance tapering are used for the left and right column in the bottom row respectively.

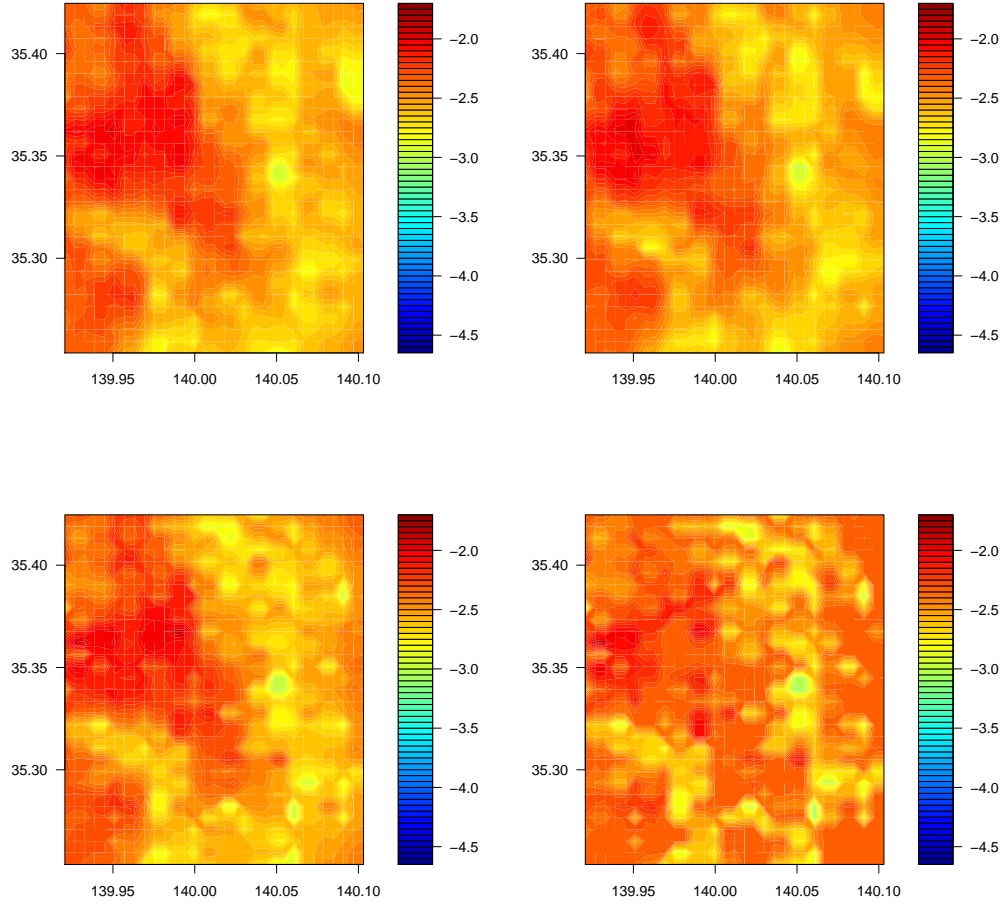


Figure 4: 95%-quantile of the predictive distribution. We adopt the stationary model and the linear projection for the left and right column in the top row respectively. The modified linear projection and the covariance tapering are used for the left and right column in the bottom row respectively.

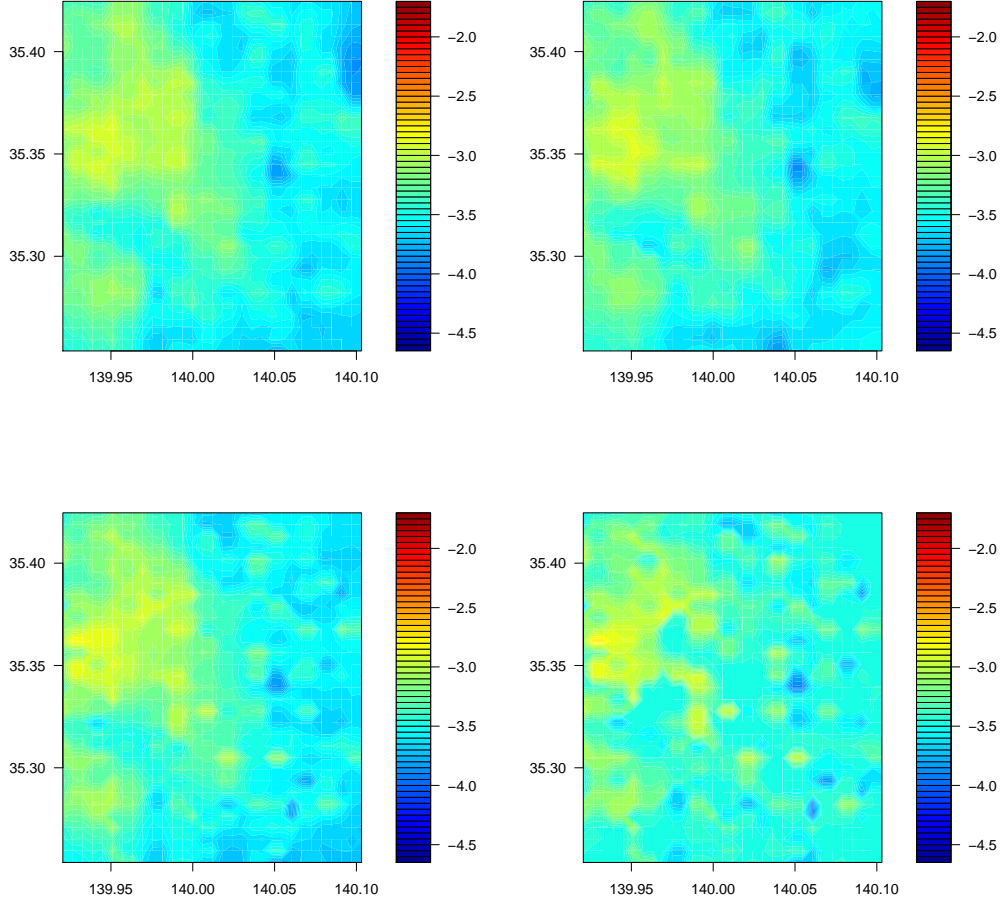


Figure 5: 5%-quantile of the predictive distribution. We adopt the stationary model and the linear projection for the left and right column in the top row respectively. The modified linear projection and the covariance tapering are used for the left and right column in the bottom row respectively.

The result of the Bayesian analysis is shown in Table 4. The modified linear projection had the lowest MSPE and DIC and outperformed even the stationary model with the exponential covariance function which induces the linear projection and modified one. This would suggest that the original spatial data set shows nonstationarity. In the four cases, the estimate of τ^2 had the relatively high value compared to that found in past empirical studies because of the large local variability which is often observed on the east side of Figure 2. Since the 95% credible intervals for β_2 in the four cases include zero, it seems that we cannot detect strong evidence of the effect of the Fukushima Dai-ichi NPP under the settings of this paper. This may be because the sampling points are not close to the Fukushima Dai-ichi

NPP and the survey was conducted less than 2 years after the TEPCO Fukushima Dai-ichi NPP Accident. A widespread sampling domain and another valid model should be used to investigate the influence of the Fukushima Dai-ichi NPP correctly.

Figure 3 is influenced by the estimate of λ from each method. The small value of λ in the linear projection results in the smooth surface. In contrast, the predictive surface of the covariance tapering has some small clusters due to the high value of λ and large local variability. The modified linear projection shares the features of both the linear projection and the covariance tapering. The original stationary model has a value of λ that is similar to the modified linear projection, but its prediction surface does not express nonstationarity. Figures 4 and 5 show large variations of the interquartile ranges at 31×31 prediction points on the east side of the sampling domain because it often has large local variability.

It is evident that our proposed modification of the linear projection using the compactly supported correlation function improves the Bayesian analysis more effectively than increasing ϵ . In addition, the modified linear projection serves as a kind of nonstationary covariance function.

7 Conclusion and future studies

In this paper, we have proposed a modified linear projection approach for huge irregularly spaced data analysis. Through some simulations and empirical study, the performance of the linear projection and covariance tapering depends on the dependence properties of the spatial covariance functions. On the other hand, our proposed method is easy to implement and is generally efficient in terms of computation time, estimation of model parameters, and prediction at unobserved locations because it effectively captures both the large- and small-scale spatial variations. Moreover, although the modified linear projection was motivated by improving the approximation of the original covariance function in the linear projection, the empirical study has shown that it can also be used as a nonstationary covariance function instead of just an approximation.

In the empirical data example, we chose the target error ϵ and taper range γ in consideration of the trade-off between prediction accuracy and computational cost for a subset of the data. In the future, we intend to develop a comprehensive selection method for these two parameters. It will also be interesting to extend the current work to non-Gaussian, multivariate, and spatio-temporal processes.

Appendix : Proofs of Propositions 1 and 2

Proof of Proposition 1

(a)

Consider $\mathbf{a}'\Phi\Sigma_W\Phi'\mathbf{a}$ for any $\mathbf{a} \in \mathbb{R}^n/\{\mathbf{0}\}$. Setting $\mathbf{b} = \Phi'\mathbf{a}$, $\mathbf{b} \neq \mathbf{0}$ because Φ is the full row-rank matrix and $\mathbf{a} \neq \mathbf{0}$. Thus, $\mathbf{b}'\Sigma_W\mathbf{b} > 0$.

(b)

Let a lower triangular matrix L be a Cholesky factor of Σ_W , that is $\Sigma_W = LL'$. Now, we have

$$\left\{ \Sigma_W - \Sigma_W\Phi'(\Phi\Sigma_W\Phi')^{-1}\Phi\Sigma_W \right\} = L \left[\mathbf{I} - (\Phi L)' \{ \Phi L (\Phi L)' \}^{-1} \Phi L \right] L'.$$

Since L is nonsingular and $\left[\mathbf{I} - (\Phi L)' \{ \Phi L (\Phi L)' \}^{-1} \Phi L \right]$ is a projection matrix from $\text{rank}(\Phi L) = m$,

$\left\{ \Sigma_W - \Sigma_W\Phi'(\Phi\Sigma_W\Phi')^{-1}\Phi\Sigma_W \right\}$ is positive semidefinite. From Theorem 5.2.1 of Horn and Johnson (1991; page 309), $\left\{ \Sigma_W - \Sigma_W\Phi'(\Phi\Sigma_W\Phi')^{-1}\Phi\Sigma_W \right\} \circ \Sigma_{taper}$ is also positive semidefinite. Therefore, $\Sigma_{sparse} + \tau^2\mathbf{I} = \left\{ \Sigma_W - \Sigma_W\Phi'(\Phi\Sigma_W\Phi')^{-1}\Phi\Sigma_W \right\} \circ \Sigma_{taper} + \tau^2\mathbf{I}$ is positive definite.

(c)

From (a), $\Sigma_W\Phi'(\Phi\Sigma_W\Phi')^{-1}\Phi\Sigma_W$ is positive semidefinite.

Thus, $\Sigma_{mlp} + \tau^2\mathbf{I} = \Sigma_W\Phi'(\Phi\Sigma_W\Phi')^{-1}\Phi\Sigma_W + \Sigma_{sparse} + \tau^2\mathbf{I}$ is positive definite.

(d)

It is clear from (a) and (b).

□

Proof of Proposition 2

In this proof, we denote $(\Sigma_W)_{ij} = a_{ij}$, $(\Sigma_{approx})_{ij} = b_{ij}$, $(\Sigma_{taper,\gamma_1})_{ij} = d_{ij}^{(1)}$, $(\Sigma_{taper,\gamma_2})_{ij} = d_{ij}^{(2)}$ and $(\mathbf{I})_{ij} = e_{ij}$ for $i, j = 1, \dots, n$. It follows that

$$\|\Sigma_W - \Sigma_{approx}\|_F = \sum_{i=1}^n \sum_{j=1}^n (a_{ij} - b_{ij})^2,$$

$$\|\Sigma_W - \{\Sigma_{approx} + (\Sigma_W - \Sigma_{approx}) \circ \mathbf{I}\}\|_F = \sum_{i=1}^n \sum_{j=1}^n (a_{ij} - b_{ij})^2 (1 - e_{ij})^2,$$

and

$$\|\Sigma_W - \{\Sigma_{approx} + (\Sigma_W - \Sigma_{approx}) \circ \Sigma_{taper,\gamma_k}\}\|_F = \sum_{i=1}^n \sum_{j=1}^n (a_{ij} - b_{ij})^2 \left(1 - d_{ij}^{(k)}\right)^2,$$

for $k = 1, 2$. From $1 \geq (1 - e_{ij})^2 \geq \left(1 - d_{ij}^{(1)}\right)^2 \geq \left(1 - d_{ij}^{(2)}\right)^2$, the proof is completed. \square

Acknowledgments The author is grateful to Professor Yasuhiro Omori and Professor Yoshihiro Yajima for their helpful comments and discussions.

References

- [1] Banerjee, A., Dunson, D. and Tokdar, S. (2013). Efficient Gaussian process regression for large data sets. *Biometrika*, **100**, 75-89.
- [2] Banerjee, S., Gelfand, A. E., Finley, A. O. and Sang, H. (2008). Gaussian predictive process models for large spatial datasets. *Journal of the Royal Statistical Society: Series B*, **70**, 825-848.
- [3] Chib, S. (2001). Markov chain Monte Carlo methods: computation and inference. In Heckman, J.J. and Leamer, E. (Eds.), *Handbook of Econometrics*, vol. 5. North-Holland, Amsterdam, 3569-3649.
- [4] Cressie, N. (1993). *Statistics for Spatial Data*, revised ed. Wiley, New York.
- [5] Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B*, **70**, 209-226.
- [6] Davis, T. A. (2006). *Direct Methods for Sparse Linear Systems*. SIAM, Philadelphia.
- [7] Finley, A.O., Sang, H., Banerjee, S. and Gelfand, A.E. (2009). Improving the performance of predictive process modeling for large datasets. *Computational Statistics & Data Analysis*, **53**, 2873-2884.
- [8] Fuentes, M. (2007). Approximate likelihood for large irregularly spaced spatial data. *Journal of the American Statistical Association*, **102**, 321-331.
- [9] Furrer, R., Genton, M. G. and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, **15**, 502-523. *Erratum and Addendum : Journal of Computational and Graphical Statistics*, **21**, 823-824.
- [10] Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D.B. (2004). *Bayesian Data Analysis*, second ed. Chapman and Hall/CRC, Boca Raton.

- [11] Gneiting, T. and Schlather, M. (2004). Stochastic models that separate fractal dimension and the Hurst effect. *SIAM Review*, **46**, 269-282.
- [12] Golub, G.H. and Van Loan, C.F. (1996). *Matrix Computations*. Johns Hopkins University Press.
- [13] Halko, N., Martinsson, P.G. and Tropp, J.A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, **53**, 217-288.
- [14] Harville, D. A. (1997). *Matrix Algebra from a Statistician's Perspective*. Springer, New York.
- [15] Hirano, T. and Yajima, Y. (2013). Covariance tapering for prediction of large spatial data sets in transformed random fields. *Annals of the Institute of Statistical Mathematics*, **65**, 913-939.
- [16] Horn, R. A. and Johnson, C. R. (1991). *Topics in Matrix Analysis*. Cambridge University Press.
- [17] Kaufman, C., Schervish, M. and Nychka, D. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, **103**, 1545-1555.
- [18] Lindgren, F., Rue, H. and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The SPDE approach (with discussion). *Journal of the Royal Statistical Society: Series B*, **73**, 423-498.
- [19] Matsuda, Y. and Yajima, Y. (2009). Fourier analysis of irregularly spaced data on R^d . *Journal of the Royal Statistical Society: Series B*, **71**, 191-217.
- [20] Paciorek, C. and Schervish, M. J. (2006). Spatial modelling using a new class of non-stationary covariance functions. *Environmetrics*, **17**, 483-506.
- [21] Plummer, M., Best, N., Cowles, K. and Vines, K. (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News*, **6**, 7-11.
- [22] Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge.

- [23] Sang, H. and Huang, J. Z. (2012). A full scale approximation of covariance functions for large spatial data sets. *Journal of the Royal Statistical Society: Series B*, **74**, 111-132.
- [24] Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society. Series B*, **64**, 583-639.
- [25] Stein, M. L. (1999). *Interpolation of Spatial Data*. Springer, New York.
- [26] Stein, M. L., Chi, Z. and Welty, L. J. (2004), Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society. Series B*, **66**:275-296.
- [27] Stewart, G.W. (1993). On the early history of the singular value decomposition. *SIAM Review*, **35**, 551-566.
- [28] Wendland, H. (1995). Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in Computational Mathematics*, **4**, 389-396.
- [29] Wikle, C. K. (2010). Hierarchical modeling with spatial data. In Gelfand, A.E., Diggle, P.J., Fuentes, M. and Guttorp, P. (Eds.), *Handbook of Spatial Statistics*. Chapman and Hall/CRC.